Corrector lingüístic informatitzat

Universitat Autònoma de Barcelona

L'ESTAT DE LA QÜESTIÓ

Abans de presentar el corrector lingüístic en què estem treballant, hem considerat apropiat fer un breu repàs de l'estat de la qüestió pel que fa a aquestes eines. També volem aclarir que partim de la idea que un corrector informatitzat ofereix alguns avantatges respecte de la consulta de diccionaris manuals en suport paper.

Els correctors lingüístics són programes que s'encarreguen de "corregir" textos i documents que es troben emmagatzemats en un fitxer d'ordinador. A partir d'un document que podem haver creat amb un tractament de textos, els correctors marquen les paraules errònies o les que són desconegudes pel corrector.

En què es diferencia el nostre projecte de corrector de la resta de correctors que hi ha al mercat?

1. El mercat ofereix bàsicament correctors associats a un programa o a una aplicació concretes. Per exemple, WordPerfect té el seu propi corrector, Word el seu, etc. Aquests correctors no poden ser utilitzats amb documents que hagin estat creats amb un altre programa que no sigui aquell per al qual han estat dissenyats. És justament en aquest aspecte que el corrector en què estem treballant es diferencia de la resta: l'actual corrector és independent de l'aplicació amb què se'l faci treballar i es pot utilitzar amb pràcticament qualsevol tractament de textos.
2. Tot i anomenar-se correctors, la majoria compleix la funció de "localitzadors" d'errors ortogràfics, i no pas de correctors estrictament. Si bé ofereixen alguns mitjans que faciliten la correcció d'un error, ha de ser el

mateix usuari qui l'haurà d'esmenar. No fan cap altre tipus d'anàlisi, no corregeixen l'estil, ni la sintaxi, sinó que es limiten a revisar l'ortografia del text (en el cas de la llengua anglesa hi ha també alguns correctors d'estil). El que fan, en realitat, és comparar cada mot del text que volem corregir amb les entrades existents al diccionari intern que cada un d'ells conté, el qual, sovint, com ara en el cas del català, acostuma a ser bastant incomplet. És a dir, cada corrector té el seu propi diccionari amb què contrasta els mots del document que estem corregint mitjançant la cerca d'una entrada que coincideixi amb la corresponent del diccionari. Si no la troba, la paraula queda marcada a la pantalla i el programa s'atura per a què decidim si volem corregir-la o bé mantenir-la.

3. Hem dit que aquests diccionaris acostumen a ser incomplets (tret, en tot cas, de l'anglès) entre altres coses perquè no utilitzen l'expansió morfològica, sinó que cada forma ha d'haver estat introduïda prèviament al diccionari (masculí, femení, singular i plural). Si alguna d'aquestes formes no existeix al diccionari del corrector però apareix al nostre text, tot i ser correcta, quedaria marcada com a paraula errònia. En canvi, si hi apareguessin les construccions "el casa", o "la cotxe", el corrector les acceptaria perquè cadascuna de les paraules és correcta per separat, encara que no concordin en el gènere.

   No cal dir que això té clars inconvenients pràctics. Quan estem utilitzant una d'aquestes eines, el corrector pot marcar paraules que siguin perfectament correctes. El motiu és que no existeixen com a entrades en el diccionari intern. En el cas de les flexions verbals és bastant usual perquè n'acostumen a contenir poquíssimes. No cal dir que com més incomplet sigui el diccionari que s'utilitza, més paraules seran marcades com a incorrectes sense que, en realitat, ho siguin. I menys òptim en serà l'ús. Al contrari, com més complet sigui, menys incidències apareixeran i més fiables seran.

4. Quan una paraula es marca com a incorrecta, aquests programes no argumenten l'error, simplement assenyalen les incorreccions. En definitiva, accepten o rebutgen un terme; si no l'accepten, ha de ser l'usuari qui ha d'esbrinar la naturalesa de l'error.

   Ara bé, tot el que hem dit fins ara no ha d'amagar el fet que són eines útils i que tenen alguns recursos que permeten suavitzar les seves limitacions. Un d'aquests recursos, que permet d'anar enriquint la pobresa inicial dels diccionaris, és la possibilitat d'incorporar al diccionari intern del corrector les paraules que són marcades com a incorrectes o inexistents si nosaltres decidim que no ho són. D'aquesta manera s'estableix una mena de diàleg i el diccionari es va ampliant amb nous termes que, si a partir d'aleshores apareguessin en un altre document, serien reconeguts com a correctes.

   També és habitual que es puguin crear diccionaris suplementaris de manera senzilla, per àmbits. Així, si estem tractant (traduint) un text amb

un llenguatge especialitzat, podem consultar el diccionari general del corrector informàtic i, simultàniament, un diccionari complementari amb entrades només de l'especialitat corresponent. Això es faria amb una única consulta, perquè el programa ja s'encarrega d'accedir a la informació de tots dos; és a dir no hem de duplicar la consulta. Això és un clar avantatge respecte dels diccionaris manuals. Un cop acabada la consulta, els dos diccionaris queden com estaven inicialment, cadascun amb les seves entrades corresponents. Per tant, el diccionari general es fa servir sempre, però podem, a més, utilitzar-ne un altre d'específic que seleccionarem segons la naturalesa de cada document. Això és, de vegades, més útil que manejar un únic diccionari molt extens perquè representa un estalvi de temps.

5. La majoria tenen la característica d'oferir un llistat de paraules com a alternativa a un mot incorrecte. Generalment, es pot seleccionar una d'aquestes paraules que serà incorporada al text que estem corregint en substitució de la incorrecta, de manera automàtica, sense que l'haguem de teclejar. Aquesta llista la fa segons uns criteris determinats:
   – Per aproximació fonètica. Però, com que acostumen a ser programes fets en anglès, els semblants que troben els fan seguint el model anglo-americà i, per tant, quan treballem amb altres llengües el resultat pot ser una mica sorprenent. Ara bé, en el cas de l'anglès, per exemple, aquesta funció pot resultar molt útil.
   – Per una certa aproximació ortogràfica. Això vol dir que si escrivim una "b" en comptes d'una "v", segurament serà capaç de trobar la paraula correcta. També ho faria en cas d'haver posat dues lletres intercanviades (amb l'ordre invertit), o en la mancança d'una lletra, una sola "s" enlloc de dues "ss", una "l" en comptes de "ll", etc. Però no gaires més sofisticacions.
6. Finalment, també és habitual que tinguin la funció de partició sil·làbica a final de línia.

Conclusió:
Les característiques essencials que tenen els correctors que podem trobar habitualment al mercat són les següents:

— Acostumen a anar associats a un programa concret.
— Localitzen errors ortogràfics i no d'una altra mena.
— Es poden enriquir els diccionaris interns i crear-ne de suplementaris.
— Poden oferir un menú de paraules alternatives que podem escollir.
— Partició sil·làbica, en molts casos.
— Es tracta d'una àrea en la que es treballa, i que ha avançat bastant des dels seus inicis. Sabem que actualment es fan investigacions que integren informació morfològica, semàntica i sintàctica, però sobretot aplicades a la llengua anglesa.

Punt de partida del projecte d'elaboració d'un corrector de textos catalans informatitzats

En el context que acabem d'exposar, podem dir que el projecte de recerca que és a la base d'aquesta comunicació, pren com a punt de partida un corrector ortogràfic concret anomenat L'ApS2 que té ben solucionat l'accés a un diccionari quant a eficiència i possibilitat d'ampliació, i que ja incorpora funcions de discriminació morfològica i algunes de sintàctiques. S'ha provat la seva eficiència entre una mostra de tres-cents usuaris.

Context general dels productes informàtics

Quan es parla d'un corrector lingüístic informatitzat s'està parlant de dues coses bastant diferents: d'un producte informàtic i d'un producte per a la llengua, ambdós com a constituents d'un sol producte. El context cultural i socioeconòmic per a un i l'altre és diferent i, en alguns trets, oposat.

La informàtica té un prestigi d'una banda bo i de l'altra dolent i ambdós repercuteixen directament sobre els seus productes. Un corrector lingüístic informatitzat pot patir-ne algunes conseqüències:

a) Haver de substituir els correctors lingüístics humans.
b) Haver de donar resposta a les peticions de correcció automàtica del text.
c) Haver d'acostumar-se a ser un producte que dóna prestigi o que simplement "fa bonic".
d) Haver de situar-se en el context lingüístic actual del català i no caure en la temptació de convertir-se en una eina que, gràcies al prestigi dels ordinadors, substitueixi l'aprenentatge i la necessitat d'enriquiment constant de la llengua.
e) Haver de respondre als requeriments per fer una traducció automàtica.
f) Haver de respondre als que li demanen una paraula "justa" i "exacta" com a alternativa a una d'incorrecta.
g) Haver de situar-se en els requeriments dels interessats que li demanen una correcció de tipus sintàctic i semàntic.

Una eina informàtica que treballi en el context de la llengua i, per tant, de la cultura ha de tenir com a un dels seus principals objectius l'enriquiment d'aquesta llengua; és a dir, ha de permetre el progressiu aprenentatge i millora de la qualitat de la llengua de l'usuari del programa, i no pas l'empobriment derivat d'haver-li facilitat les coses més del compte. Ha d'oferir-li alternatives i suggeriments, facilitar-li l'elaboració de les seves pròpies decisions, donar-li informació i ajudar-lo a enriquir les seves pròpies opcions. Ha d'adaptar-se al nivell de l'usuari sense forçar-lo a uns constrenyiments preestablerts que seran en uns casos massa còmodes i en altres insuficients.

L'ordinador, en aquest context i com a eina d'ajuda, pot fer créixer les possibilitats d'aprofundiment i expansió de la comunicació entre les llengües i, en conseqüència, entre les cultures. Treballar pel desenvolupament d'eines que facilitin l'ús i coneixement de les diferents llengües no pot fer res més que afavorir la comunicació, l'enriquiment cultural i el respecte mutu entre els pobles.

El corrector L'ApS2

L'equip que treballa en el projecte d'elaboració d'un corrector lingüístic en català està format per quatre persones, dos lingüistes i dos informàtics.

L'ApS2 és un corrector de textos per al català, que treballa independentment dels processadors de textos, és a dir, que s'adapta a la majoria de processadors i a l'entorn de treball. L'ApS2 consta d'un diccionari principal que conté, amb expansió morfològica inclosa, uns 300.000 mots (això fa un total d'uns 50.000 mots sense expansió morfològica). L'ApS2 aparegué al carrer a començaments del 1990 i ja llavors incorporava per al català les funcions d'ortografia que després han anat incorporant per a altres llengües tots els processadors que volen abastar quotes amples del mercat. Aquest corrector conté bàsicament l'accés a un diccionari català en el qual busca totes les paraules problemàtiques. En cas de no trobar-les-hi, n'aportarà de semblants amb l'objectiu de trobar el mot correcte. Aquest primer algorisme que troba les paraules semblants s'ha fet a partir de les regles fonètiques i ortogràfiques catalanes i també a partir d'estudis lingüístics diversos (per a l'anglès, és clar...). L'ApS2 s'ha beneficiat d'estudis sobre fonètica realitzats amb gent que pateix disfuncions de la parla. De resultes d'aquests estudis sortiren unes regles de proximitat fonètica i ortogràfica entre les paraules que, posteriorment, s'empobririen per tal d'aconseguir eficiència de resposta en els primitius PCs i XTs. Aquestes regles s'ampliaran i es reconsideraran de nou en el context del projecte. També s'analitzarà la inclusió, o no, de criteris de detecció d'errors mecanogràfics que una versió anterior de L'ApS2 contingué però que ara es revisarà.

Les funcions anteriors s'insereixen, doncs, en el context de les funcionalitats normals dels correctors ortogràfics inclosos en els processadors de textos actuals. També permet de trobar la descomposició sil·làbica d'un mot qualsevol o de crear diccionaris personals que ampliïn el diccionari principal que incorpora el programa.

L'ApS2, a més, té un algorisme d'expansió morfològica de les formes verbals basat en l'infinitiu i en el model de conjugació. En el context del projecte possiblement s'ampliï l'algorisme i la base lingüística a noves anàlisis i a altres tipus d'expansió morfològica. L'algorisme d'expansió morfològica de les formes verbals del català és prou eficaç per a permetre que l'accés

al diccionari sigui igual de ràpid que el dels programes processadors de textos més usuals del mercat, tot i que no conté una funció com aquesta.

L'ApS2 detecta, també, les contraccions de pronoms i de morfemes que s'han fet malament o que no s'han fet quan s'havien de fer. Evidentment aquesta és una funció específica per al català i que, per la seva dificultat d'elaboració, no és fàcil que la incorporin gaires correctors ortogràfics en un futur proper.

Per fi, el diccionari de L'ApS2 no és únicament lexicogràfic ja que també conté un recull de formes col·loquials, barbarismes, possibles errades, etc. En aquests casos afegeix un curt comentari explicatiu, exemple aclaridor, avís sobre possibles descuits dels accents, etc. Actualment hi ha prop de 1.600 mots en aquestes condicions i, com a part del projecte, es revisaran, es classificaran, s'ampliaran i s'enriquiran.

Millora i ampliació de les prestacions dels corrector L'ApS2

El projecte de millora i d'ampliació de l'actual corrector L'ApS2 ha obtingut una subvenció de la CIRIT dins dels projectes considerats d'"Alta Qualitat". Aquest projecte, però, forma part del que nosaltres anomenem primera fase, pel fet que L'ApS2 és susceptible de ser un corrector aplicable a entorns diferents a l'entorn DOS.

Bàsicament, el projecte d'ampliació vol millorar tant les prestacions informàtiques com les lingüístiques. No cal dir que en aquest projecte els encarregats de la part lingüística hauran de treballar estreta i conjuntament amb els de la part informàtica. Les principals tasques a realitzar són les següents:

1. Ampliar el diccionari intern, que per ara, com ja s'ha dit, consta d'unes 300.000 entrades, comptant les expansions morfològiques. Ara es vol incorporar el Diccionari de l'Enciclopèdia Catalana i totes les expansions morfològiques dels verbs.
2. Revisar i ampliar els comentaris associats a les possibles incorreccions que es produeixin en el text que es vol corregir. Cada error, per tant, conté informació gramatical sobre l'error comès i ajuda l'usuari a seleccionar la resposta correcta. En alguns casos d'ambigüitat o d'errors comuns, el comentari associat pot arribar a fer la funció d'una petita gramàtica aclaridora de dubtes.
3. Ampliar el llistat de barbarismes. L'ApS2 ja conté un llistat dels més freqüents que arribarà a incloure, en aquesta primera fase, 500 vocables. La correcció del barbarisme contemplarà els diferents dialectes catalans.
4. La contemplació dels diferents dialectes afectarà també els mots ja existents en el diccionari. En aquesta primera fase es contempla l'entrada de paraules pertanyents al dialecte balear i al valencià.

5. Ampliació del llistat de sinònims, que per ara és encara força limitat.
6. En una darrera fase es vol que l'ApS2 pugui ser un corrector sintàctic i estilístic. Per això s'establirà un mètode que permeti de detectar les combinacions pronominals correctes i les incorrectes i suggerir possibles solucions per a l'error. També podrà detectar la repetició de mots en línies properes per tal de fer veure a la persona que corregeix el text si la repetició és volguda o involuntària. En cas que es vulgui esmenar la repetició, el corrector suggerirà possibles mots alternatius mitjançant l'activació del llistat de sinònims. També podrà detectar casos de pleonasme o, fins i tot, la manca de pronominalització.
7. Actualment, L'ApS2 marca les incorreccions i dóna la possible correcció mitjançant un sistema de semblança. Davant d'una paraula incorrecta com "camibar" troba possibles correccions segons un criteri d'aproximació morfològica; per tant, és fàcil que suggereixi la correcció "caminar". Amb tot, el criteri de semblances no és encara prou acurat, és massa ampli, de manera que les possibles correccions s'allunyen molt de la paraula correcta. Caldrà, doncs, establir un criteri informàtic que aproximi l'errada a la trobada d'aquest mot. Per altra banda, lingüísticament s'estudiaran els casos d'errors freqüents semblants als errors dislèxics, "ces" per "sec", per exemple, o els casos freqüents d'errors de teclejat.
8. Finalment, s'introduirà la funció de partició sil·làbica que encara cap corrector no té adaptada al català.